Estimating the body portion of CT volumes by matching histograms of visual words

Johannes Feulner[a], S. Kevin Zhou[b], Sascha Seifert[c], Alexander Cavallaro[d], Joachim Hornegger[a] and Dorin Comaniciu[b]

[a]Chair of Pattern Recognition, University of Erlangen, Martensstrasse 3, 91058 Erlangen, Germany;
[b]Siemens Corporate Research, Inc., 755 College Road East, Princeton, NJ 08540;
[c]Siemens Corporate Technology, Günther-Scharowsky-Str. 1, 91058 Erlangen, Germany
[d]Imaging Science Institute Erlangen, Maximiliansplatz 1, 91054 Erlangen, Germany

ABSTRACT

Being able to automatically determine which portion of the human body is shown by a CT volume image offers various possibilities like automatic labeling of images or initializing subsequent image analysis algorithms. This paper presents a method that takes a CT volume as input and outputs the vertical body coordinates of its top and bottom slice in a normalized coordinate system whose origin and unit length are determined by anatomical landmarks. Each slice of a volume is described by a histogram of visual words: Feature vectors consisting of an intensity histogram and a SURF descriptor are first computed on a regular grid and then classified into the closest visual words to form a histogram. The vocabulary of visual words is a quantization of the feature space by offline clustering a large number of feature vectors from prototype volumes into visual words (or cluster centers) via the K-Means algorithm. For a set of prototype volumes whose body coordinates are known the slice descriptions are computed in advance. The body coordinates of a test volume are computed by a 1D rigid registration of the test volume with the prototype volumes in axial direction. The similarity of two slices is measured by comparing their histograms of visual words. Cross validation on a dataset of 44 volumes proved the robustness of the results. Even for test volumes of ca. 20cm height, the average error was 15.8mm.

Keywords: Body portion estimation, inter subject registration, bag of features, visual words

1. INTRODUCTION

This paper addresses the problem of determining which portion of the body is shown by a stack of axial CT image slices. For example, given a small stack of slices containing the heart region, one may want to automatically determine where in the human body it belongs.

This offers various applications like attaching text labels to images of a database. A user may then search the database for volumes showing the heart. The DICOM protocol already specifies a flag "Body part examined", but this is imprecise as it only distinguishes 25 body parts, and is even wrong in many cases as reported by Gueld et al.[1] Or alternatively, it may be used to reduce traffic load on medical image databases: Often physicians are only interested in a small portion of a large volume stored in the database. If it is known which parts of the body the large image shows, the images slices of interest showing e.g. the heart can be approximately determined and transferred to the user. Another possible application is pruning the search space of subsequent image analysis algorithms like organ detectors.

Further author information: (Send correspondence to J.F.)
J.F.: E-mail: johannes.feulner@informatik.uni-erlangen.de, Telephone: +49 9131 8527825
S.K.Z.: E-mail: shaohua.zhou@siemens.com, Telephone: +1 609 7343325
S.S.: E-mail: saschaseifert@siemens.com, Telephone: +49 9131 731392
A.C.: E-Mail: alexander.cavallaro@uk-erlangen.de, Telephone: +49 9131 8545515
J.H.: E-mail: joachim.hornegger@informatik.uni-erlangen.de, Telephone: +49 9131 8527883
D.C.: E-mail: dorin.comaniciu@siemens.com, Telephone: +1 609 7343643

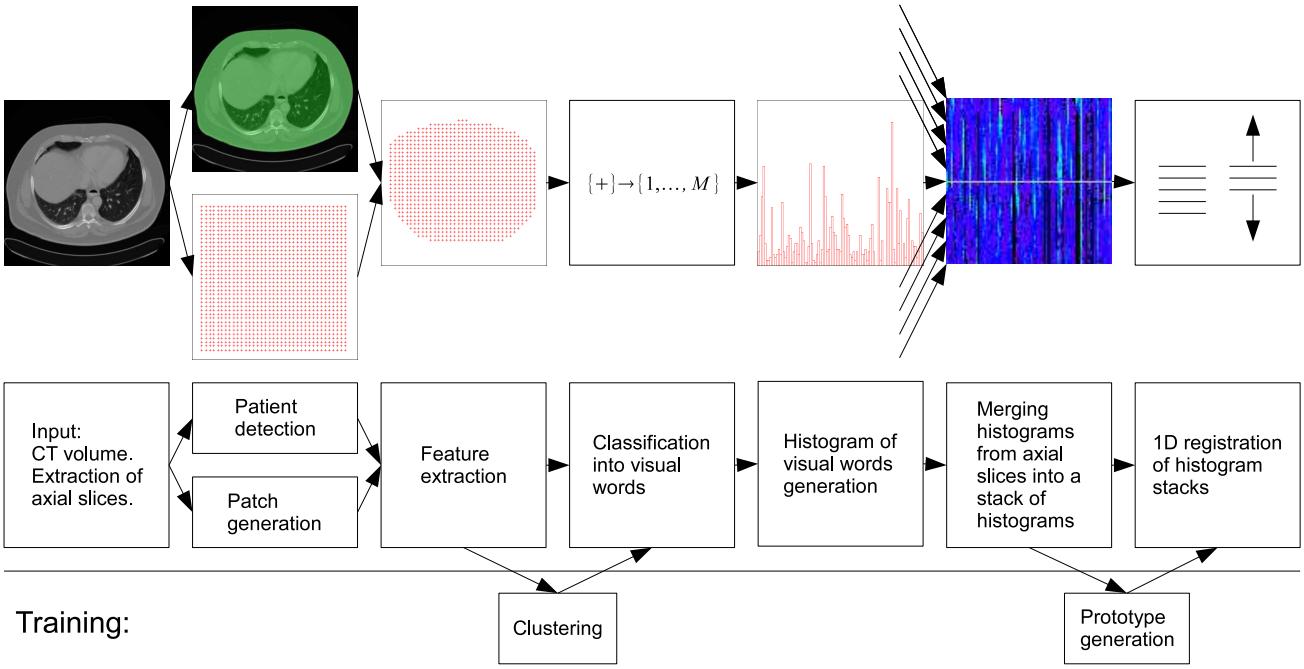

Figure 1. The proposed system for body portion estimation. The axial slices of a CT volume are first processed separately. sample positions are generated on a regular grid. For each sample position inside the patient, a SURF descriptor is computed from the local neighborhood called "patch". The descriptors are classified into visual words and accumulated in a histogram. The stack of histograms from the axial slices is registered with prototype histogram stacks to find the body portion.

The problem of estimating the body portion is closely related to inter-subject image registration as it can be solved by registering the volume to an anatomical atlas. This is typically solved in two ways: By detecting anatomical landmarks in the volume image, or by intensity based non-rigid image registration. Landmark based registration may also be used as an initialization for non-rigid registration. However, a set of landmark is required that covers all regions of the body and can be robustly detected. Intensity based registration tends to be slow, and because it is prone to getting stuck in local optima, it requires a good initialization. In many cases one is only interested in registration along the longitudinal (z) axis and a complete 3D registration is not necessary.

Dicken et al.[2] proposed a method for recognition of body parts covered by CT volumes. An axial slice is described by a Hounsfield histogram with bins adapted to the attenuation coefficient of certain organs. Derived values like the spatial variance within the slice of voxels of a certain bin are also included into the descriptor. The stack of the N-dimensional axial slice descriptors is interpreted as a set of N 1D functions whose domain is the (vertical) z level. Then five handcrafted rules are used to decide which body parts are visible, where eight different body parts are distinguished. However, the results are imprecise because no quantitative estimation of the covered body region is performed. Furthermore, they report problems with short scan ranges.

For the purpose of scene classification, is has recently become popular to measure the similarity of two images by extracting a bag of features from both images. Grauman and Darrell[3] proposed a distance measure for feature bags which builds a pyramid of histograms of features and compares two histogram pyramids. Lazebnik[4] adapted this distance measure by first classifying the feature vectors into visual words. The vocabulary is generated in advance by clustering feature vectors extracted from a set of training images. Then a spatial pyramid of histograms of the visual words is generated and used to compare two images.

In this paper, histograms of visual words are used to register stacks of CT image slices. Only the z axis of the volume is considered as it is sufficient for many applications and leads to a small search space that even allows exhaustive search. By 1D registration of a test volume along the longitudinal axis to prototype volumes whose body region is known, the body region of the test volume is estimated. In order to measure body regions,

1D "body coordinates" (bc) are introduced, whose origin level is defined to be a landmark in the pelvis, and the unit length is chosen to be the distance between a landmark at the clavicle and the pelvis landmark.

Figure 1 shows an overview of the proposed system. For an incoming volume, first the skin of the patient is detected. Independently of this, the axial slices of the volume are regularly divided into small quadratic patches. In the next step, a feature vector is extracted from each patch, which is used to classify the patch into a visual word belonging to a predefined vocabulary. Only patches inside the skin of the patient are considered in order not to get confused by the environment, e.g. the table the patient lies on and the air surrounding the patient. A visual word corresponds to a class of patches sharing similar appearances. Now a histogram is generated from the visual words detected in a slice of the volume, which serves as a description of the slice. With this being performed for all slices of the volume, the result is a stack of histograms. A set of training volumes with known annotations of the pelvis and clavicle landmarks are processed in the same way, resulting in a set of prototype histogram stacks. The vocabulary of visual words is generated in advance by clustering the feature vectors extracted from the training volumes. Now the body portion of the input volume is determined by 1D registration of its histogram stack with respect to the prototype stacks with known body regions. Generally a single prototype would be enough, but using more than one leads to more robust results.

The structure of the rest of this document is as follows: In section 2 the extraction of visual words, the histogram generation and the features used are explained. Section 3 is on the registration of the histogram stacks. Section 4 describes experiments and presents results, and section 5 concludes the paper.

2. HISTOGRAMS OF VISUAL WORDS

Visual words are primitive patches used to characterize an ensemble of images. The visual word vocabulary may include straight lines, corners, uniform patches, holes or certain textures.

The concept of describing images using a visual vocabulary has been successfully used before for purposes of data mining, scene classification and object recognition. Bhattacharya et al.[5] described retina images using a visual vocabulary. This description was used to distinguish image classes and to highlight parts of the image that are characteristic for their class. Duygulu et al.[6] labeled image regions with keywords from a predefined vocabulary of nouns in order to automatically generate an image description and to recognize objects.

In this paper histograms of visual words are used to measure the similarity of two axial CT slices. Visual words are extracted from an image slice on a regular grid. For a typical slice of size 35cm by 35cm, about 1000 samples or visual words are extracted.

A simple patient detector is used to reject samples outside the patient. The skin of the patient is detected by scanning rows and columns from both directions until a certain number of pixels is above a threshold of -600 HU. This proved to be fast and effective for rejecting the air surrounding the patient and also the table s/he lies on.

For all sample points inside the patient, a 72 dimensional feature vector is computed, which consists of an eight bin histogram of the Hounsfield units and a 64 dimensional oriented SURF descriptor.[7] SURF descriptors were used because good results were reported for scene classification with SIFT features.[4,8,9] SURF approximates SIFT but is faster to compute as it uses Haar-like filters and integral images to speed up computation. Refer to Ref. 7 for details about SURF descriptors.

As SURF descriptors were designed to be invariant to illumination changes that often cause problems in computer vision, they do not make use of absolute intensities. However, in CT images absolute intensities are reliable. In order to use this information, the SURF descriptor is extended with the Hounsfield histogram, which is scaled to fit the mean values of the SURF descriptor entries. Descriptors are computed at a fixed scale of 2.5, which corresponds to a descriptor window size of 50×50 pixel. An alternative to a fixed sampling grid and fixed scaling is to detect key locations in the image, for example minima and maxima in scale space as suggested by Lowe,[9] but better results were reported for a regular dense sampling by Fei-Fei and Perona.[8]

The extracted feature vectors are now classified into a set of visual words. The vocabulary is represented by a prototype feature vector for each word, and for classification the nearest neighbor is used. The distance of two feature vectors is measured using the ℓ^2 norm. To generate the vocabulary, a random subset of feature vectors is

extracted from a set of training images, and the K-Means algorithm is used to find clusters. The cluster centers are chosen as the vocabulary. In Figure 2, one example image from each visual word cluster is displayed. In Figure 3, a stack of histograms of visual words is shown together with a coronal section of the original volume.

For each slice, a histogram of visual words is generated. This serves as a description, which is used to measure similarity between slices. This is similar to the method described by Lazebnik[4] but does not make use of a spatial pyramid in order not to make any assumption of the patient position, which is usually in supine position, but can also be in prone position or lying on the side. Also, the patient is not necessarily centered in an image slice.

3. HISTOGRAM MATCHING

To measure the distance d of two slices s and t, their histograms H_s and H_t are compared using the sum of absolute differences (SAD)

$$d(s,t) = \sum_{i=0}^{M-1} |H_s(i) - H_t(i)|. \tag{1}$$

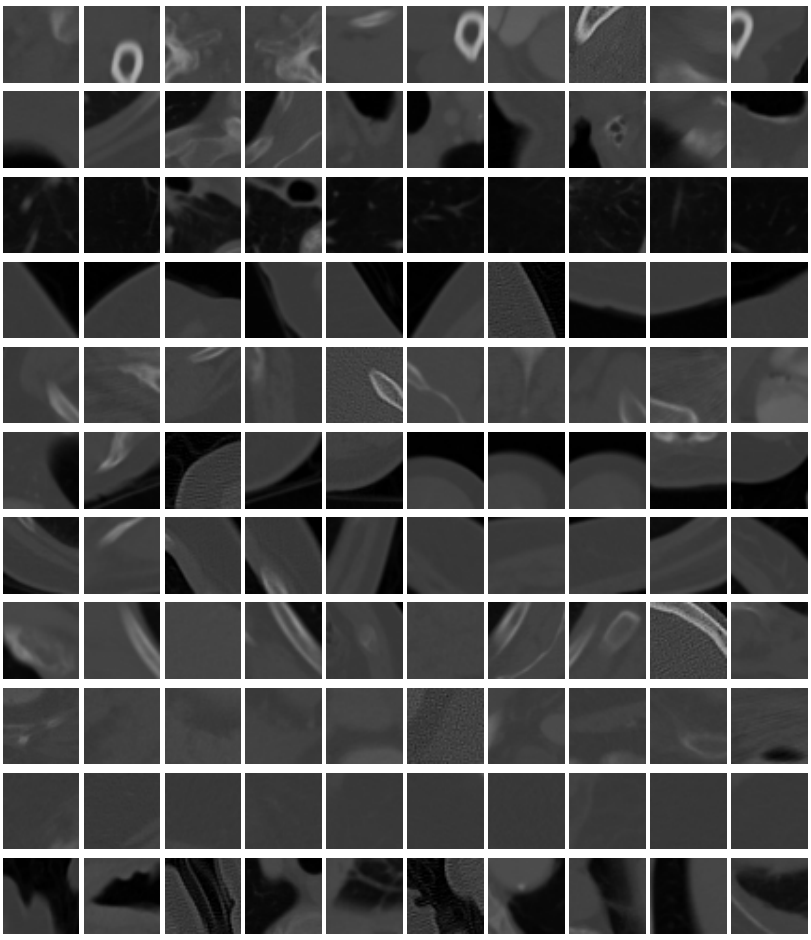

Figure 2. Example images are shown for ten different visual words picked from a vocabulary of size 100. A row in the image corresponds to a visual word. Some correspond to homogeneous regions at a certain attenuation coefficient, other to air-soft-tissue edges, soft-tissue-bone edges, straight or curved edges, or holes/blobs.

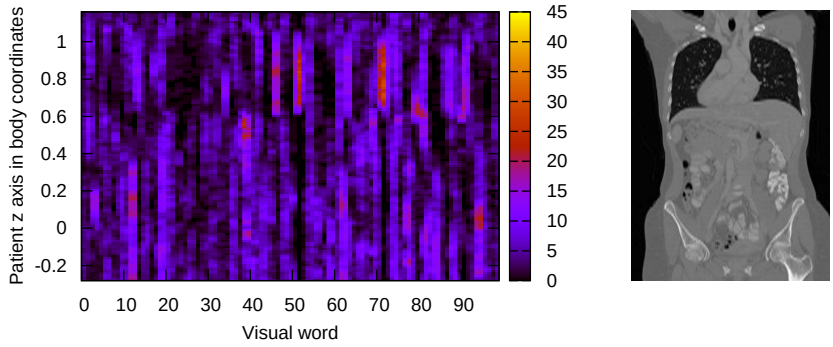

Figure 3. Histograms of visual words along with a coronal section of the volume it was generated from. Salient are especially the visual words that correspond to the lung region. The image is best viewed in color.

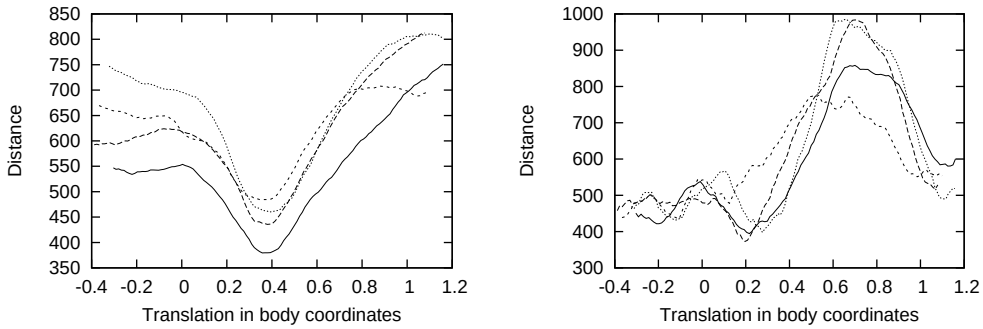

Figure 4. Objective function f shown for four different prototype volumes. Left: Test volume with 114 slices. Right: Test volume with 10 slices from the abdomen. For the large volume, one clear minimum exists. For the small stack it is more ambiguous. But still in 3 out of 4 cases, the global optimum is close to the correct location (at approx. 0.2bc).

which is for a fixed number of visual words per image up to a normalizing factor equivalent to using one minus the histogram intersection[10] h

$$h(s,t) = \sum_{i=0}^{M-1} \min(H_s(i), H_t(i)). \tag{2}$$

Here, M denotes the number of histogram bins that equals the size of the vocabulary. The objective function f used to rigidly register two slice stacks $S = s_0, \ldots, s_{n-1}$ and $T = t_0, \ldots, t_{m-1}$ along the z axis is the average distance of their slices

$$f(z) = \frac{1}{i_{max} - i_{min} + 1} \sum_{i_{min}}^{i_{max}} d(t_i, s_{i+z}), \tag{3}$$

where i_{max} and i_{min} are chosen so that there is at least 50% overlap between the two stacks S and T. For the z axis a discretization of 5mm was chosen. Because a single evaluation of the objective function f is computationally inexpensive and the search space is only one-dimensional, exhaustive optimization is feasible. Figure 4 shows $f(z)$ for two test stacks $T_{1,2}$ of different size and four reference histogram stacks $S_{1 \ldots 4}$.

After exhaustive optimization, a set of candidates $C = \{c_1, c_2, \ldots, c_{||C||}\}$ is generated from f by finding local optima. The reason is that especially for volumes with a small number of slices, it occasionally happens that the global optimum is not the right solution. However, the correct solution is almost ever located in a valley. A weight w_i is now attached to each candidate c_i, which is computed from the objective function at c_i and its

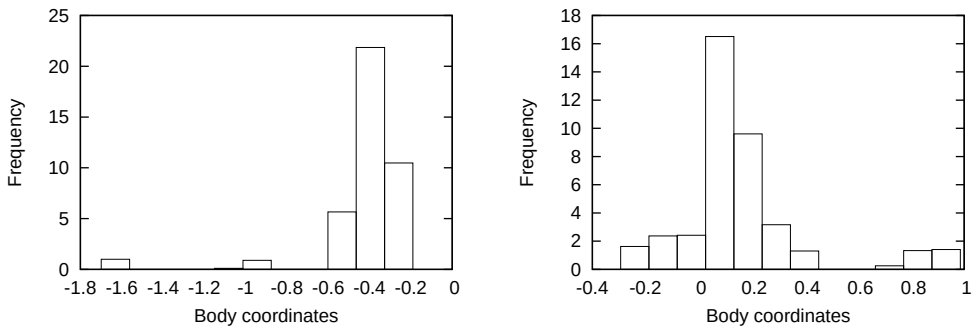

Figure 5. Histogram of location candidates, with one from each prototype. 40 prototypes were used in total. Left: Test volume with 114 slices. Right: Test volume with 10 slices from the abdomen. For the small test volume, the variance is higher, but still a single obvious mode exists.

Num. partitions/ size in mm*	10/44	5/86	3/140	2/206	1/427	Total
e in mm	45.3 ± 77.7	31.7 ± 59.6	21.2 ± 33.9	15.8 ± 16.9	16.6 ± 12.8	34.4 ± 63.6
$e < 10$mm in %	31.1	38.2	47.7	44.3	36.4	36.7
$e < 20$mm in %	51.1	59.5	75	76.1	72.7	60
$e < 30$mm in %	65.9	74.1	84.1	86.4	86.4	73.4
$e < 40$mm in %	73	79.1	88.6	96.6	93.2	79.9
$e < 50$mm in %	79.3	85.9	90.9	98.9	95.5	85.2
$e < 60$mm in %	82.5	90	93.2	98.9	100	88.2
$e < 100$mm in %	90.2	94.5	95.5	98.9	100	93.3
e_{HU} in mm	71.4 ± 100	60 ± 96.6	54.7 ± 85.1	40.5 ± 71.1	51.3 ± 89.4	62.4 ± 94.8
$e_{\mathrm{HU}} < 60$mm in %	67.7	76.8	76.5	86.4	88.6	73.9
Num. registrations	440	220	132	88	44	924

Table 1. Results of registration accuracy and robustness. The columns show results for test volumes of different height, from 4cm to 43cm. First row: mean registration error along with standard deviation in millimeters. Below: Percentage of cases where registration error e was better than a threshold. e_{HU}: Accuracy of a registration only based on histograms of attenuation coefficients for comparison. Bottom row: Number of registration this column was generated from. *Size of partition in mm is an approximate value, averaged over patients.

second derivatives:

$$w_i = 2\left(\sum_{j=0}^{2} f(z-c_i) * g_j(z)\right) - f(c_i). \tag{4}$$

Here, $*$ denotes convolution, g_0 is a filter kernel to compute the second derivative, and $g_{j+1}(z) = g_j(\frac{z}{3})$ is scaled with a factor of 3 relative to g_j.

In order to achieve robust results, a test volume is registered with several prototype volumes. To select the final candidate, first for each single registration the candidate with the best weight is selected, resulting in a set of best candidates $B = \{b_1, \ldots, b_{||B||}\}$. To become insensitive to single candidates with good weights far away from most others, the mode z_m in a histogram is detected and the final result is the candidate with the best weight closer to z_m than a certain threshold θ. In Figure 5, two examples for histograms of B are shown.

Note that, though the described method does not handle scale variations explicitly as they occur for patients of different size, they are handled implicitly by the scale variations of the training data. For instance, a test volume showing a tall patients will generally fit best to tall patient in the training set.

4. RESULTS

Registration accuracy was evaluated using 44 CT volume scans showing the thoractic and abdominal region. For all datasets, annotation of landmarks at the clavicle and the pelvis were available. They served as ground truth

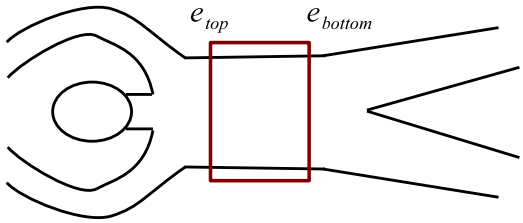

Figure 6. Illustration of the error measure used. For a single registration, the error was measured at the top and bottom of the test volume.

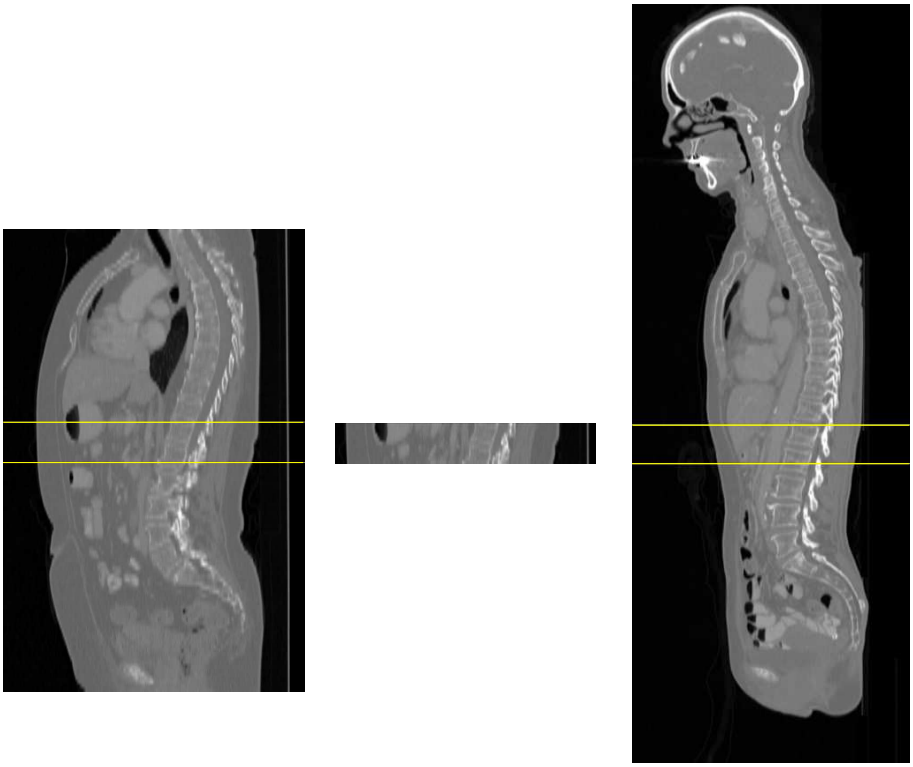

Figure 7. Example of a registration result. Middle: Sagittal slice through the test sub volume of which the body region is to be determined. It consists of 10 axial slices with a slice thickness of 5mm and shows a portion of the abdomen. Left: True position in the original volume from which the sub volume was cropped. Right: Sagittal slice through a volume with known body coordinates. The horizontal lines show the estimated body region covered by the test sub volume.

for the body coordinate system, marking the levels zero and one. In between, linear interpolation was used to generate ground truth values for the body coordinates.

Eleven fold cross validation was used to separate the datasets in test and prototype volumes. Registration was performed with slice stacks of five different sizes: A test stack was always partitioned into ten, five, three, two and one pieces, resulting in $10 + 5 + 3 + 2 + 1 = 21$ registrations per fold and test volume.

The error of a single registration was measured at the top and the bottom of the test volume (see Figure 6 for illustration). The average of the absolute values

$$e = \frac{1}{2}\left(|e_{\text{top}}| + |e_{\text{bottom}}|\right) \tag{5}$$

was taken as the final error e. Table 1 shows the results of the cross evaluation. The columns show the registration accuracy for the five different test volume heights. Generally, accuracy improves with the height of the test volume because more context can be used. For the test volumes with the smallest size (a height of 4.4cm), the average error is about the same as the size. For mid-size test volumes of 21cm height, the average error was 15.8mm,

and in 87 out of 88 cases, the error was below 5cm. For 43cm test volumes, estimation worked in all 44 cases with a maximum error of less than 6cm with an average error of 16.6mm and a standard deviation of 12.8mm. For comparison, registration was also evaluated using a 1024-bin histogram of the Hounsfield units as a slice descriptor. Inter-slice-similarity was again measured using the sum of absolute differences of the histograms. The histogram of visual words clearly outperformed the Hounsfield histogram (see e_{HU} in Table 1).

Figure 7 shows an example of the algorithm's output. The input is a portion of the abdomen of 10cm height. To visualize the result, another volume shown at the right side was annotated with body coordinates. The horizontal lines at the right indicate the estimated body region. The horizontal lines at the left show the true position in the original volume.

As the proposed algorithm is deterministic, its computation time was only benchmarked on a single dataset of 100 slices and using 40 prototype volumes. On a PC with Intel Core 2 Duo 2.2GHz CPU and 2GB of RAM, patient detection took 0.13s, computing the histogram of visual words 15.2s, and exhaustive search 0.77s. The algorithm can be parallelized easily. We leave this for future work.

5. CONCLUSION

This paper presents a method for estimating the body region of a CT volume image. It is based on 1D registration of histograms of visual words, which serve as a description of a CT slice. Experiments showed that especially for mid-sized and large volumes the body region can be estimated very robustly with an mean error of approximately 2cm. Besides automatic initialization of further processing steps like organ detection, possible applications are also automatic labelling of images for the purpose of semantic image search.

REFERENCES

[1] Gueld, M. O., Kohnen, M., Keysers, D., Schubert, H., Wein, B. B., Bredno, J., and Lehmann, T. M., "Quality of dicom header information for image categorization," *Medical Imaging 2002: PACS and Integrated Medical Information Systems: Design and Evaluation* **4685**(1), 280–287, SPIE (2002).

[2] Dicken, V., Lindow, B., Bornemann, L., Drexl, J., Nikoubashman, A., and Peitgen, H.-O., "Realtime image recognition of body parts scanned in computed tomography datasets," *22nd International Congress on Computer Assisted Radiology and Surgery – CARS 2008* (2008).

[3] Grauman, K. and Darrell, T., "The pyramid match kernel: discriminative classification with sets of image features," *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on* **2**, 1458–1465 Vol. 2 (Oct. 2005).

[4] Lazebnik, S., Schmid, C., and Ponce, J., "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* **2**, 2169–2178 (2006).

[5] Bhattacharya, A., Ljosa, V., Pan, J.-Y., Verardo, M. R., Yang, H., Faloutsos, C., and Singh, A. K., "Vivo: Visual vocabulary construction for mining biomedical images," in [*ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*], 50–57, IEEE Computer Society, Washington, DC, USA (2005).

[6] Duygulu, P., Barnard, K., de Freitas, J. F. G., and Forsyth, D. A., "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in [*ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part IV*], 97–112, Springer-Verlag, London, UK (2002).

[7] Bay, H., Tuytelaars, T., and Van Gool, L., "Surf: Speeded up robust features," *Computer Vision ECCV 2006* , 404–417 (2006).

[8] Fei-Fei, L. and Perona, P., "A bayesian hierarchical model for learning natural scene categories," *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* **2**, 524–531 vol. 2 (June 2005).

[9] Lowe, D., "Object recognition from local scale-invariant features," *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on* **2**, 1150–1157 vol.2 (1999).

[10] Swain, M. J. and Ballard, D. H., "Color indexing," *Int. J. Comput. Vision* **7**(1), 11–32 (1991).